JESLIB 2012; 1(1): 4-17
doi:10.7191/jeslib.2012.1004

Journal of eScience Librarianship
putting the pieces together: theory and practice

DataONE: Facilitating eScience Through Collaboration

Suzie Allard

University of Tennessee, Knoxville, TN, USA

Abstract

Objective: To introduce DataONE, a multi-institutional, multinational, and interdisciplinary collaboration that is developing the cyberinfrastructure and organizational structure to support the full information lifecycle of biological, ecological, and environmental data and tools to be used by researchers, educators, and the public at large.

Setting: The dynamic world of data intensive science at the point it interacts with the grand challenges facing environmental sciences.

Methods: Briefly discuss science's "fourth paradigm," then introduce how DataONE is being developed to answer the challenges presented by this new environment. Sociocultural perspectives are the primary focus of the discussion.

Results: DataONE is highly collaborative. This is a result of its cyberinfrastructure architecture, its interdisciplinary nature, and its organizational diversity. The organizational structure of an agile management team, diverse leadership team, and productive working groups provides for a successful collaborative environment where substantial contributions to the DataONE mission have been made by a large number of people.

Conclusions: Librarians and information science researchers are key partners in the development of DataONE. These roles are likely to grow as more scientists engage data at all points of the data lifecycle.

Introduction

EScience is changing the way librarians work and the services they provide. An important aspect of eScience is the focus on data, as noted by Kafel (2010): "A prominent feature of eScience is the generation of immense data sets that can be rapidly disseminated to other researchers via the internet." There is an enormous increase in the amount of data collected, analyzed, re-analyzed, and stored, which is a result of developments in computational simulation and modeling, automated data acquisition, and communication technologies (National Academies of Science 2009). These data-intensive activities present challenges that librarians will be addressing with their science communities and that librarians are uniquely trained to negotiate successfully. Beyond technological changes, as scientific research is becoming more data intensive, a "fourth paradigm" (Hey, Tansley, and Tolle 2009) has emerged. Gray (2007) identifies the first three paradigms over a temporal span beginning at a thousand years ago when science was empirically describing natural phenomena. In the last few hundred years, science added a theoretical branch using models and generalizations. Within the last few decades, science added a third paradigm which is a computational branch enabling simulations. The fourth paradigm is emerging now and is best described as data

Correspondence to Suzie Allard: sallard@utk.edu
Keywords: eScience, DataONE, data-intensive science, cyberinfrastructure

exploration that unifies theory, experiment, and simulation. It is often referred to as eScience. The fourth paradigm is changing how science is conducted (Hunt, Baldocchi and van Ingen 2009), as well as how scientists and publishers engage the scholarly record (Lynch 2009). The fourth paradigm, eScience, focuses on unifying theory, experiment, and simulation. The sociocultural changes brought about by the fourth paradigm also have implications for libraries and librarianship, suggesting the extension of current relationships within the scientific community, including publishers and the development of new collaborations. Ultimately, the key to benefitting society is to find solutions to the challenges that arise from conducting data intensive science (Hey, Tansley and Tolle 2009).

The science librarian can play an essential part in enabling the cyberinfrastructure, including both technology and people, that supports eScience, but this role is still emerging and may not be adequately defined in existing job descriptions. This paper is designed to help set the context of eScience so that the role of the eScience librarian can be explored. The paper begins by briefly discussing the cyberinfrastructure that is needed to make eScience successful, and then introduces one project, DataONE, as an exemplar to illustrate how a cyberinfrastructure may be configured, with particular attention to the participation of librarians.

The Need for Cyberinfrastructure

Many scientific problems are both data intensive and complex. For example, the grand challenges facing science, such as climate change (International Panel on Climate Change 2007), destructive pandemics (World Health Organization 2009), or sustainable energy (World Energy Council 2010), are not confined to one or two disciplines, but rather cross many scientific domains, creating a situation in which the information is becoming more interconnected (Hannay 2009). Recognizing that interconnections exist is important because it allows us to address complex issues with a better contextual understanding. However, interconnected information demands that we be able to make sense of information across disparate vocabularies, heterogeneous information artifacts, and diverse paradigms. This creates intellectual and technological challenges that may not be addressed sufficiently with traditional information tools and methods. It also suggests new roles for the information managers and librarians who work with the information, and for the people who create and use the information.

The foundation to successfully negotiate this complex data intensive environment is a robust cyberinfrastructure that provides the technology and associated tools to support scientists in their activities and to facilitate new ways to engage science (National Science Foundation Cyberinfrastructure Council 2007). The definition of cyberinfrastructure includes technological and sociological perspectives (National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure 2003). Both perspectives are needed to address the challenges presented by the increased amount of data collected, analyzed, and stored, including a heightened need for technology that assures data preservation, for processes that enable digital curation, and for approaches to enable metadata interoperability. This means that data intensive science challenges extend beyond the traditional hard sciences and require research engagement from the social sciences. It also suggests that while data-driven science requires persistent and reliable data and tools for scientists to create and use these data, it also will benefit from tools that can be used by a variety of stakeholders beyond scientists, including government decision-makers, academic researchers, industry leaders, non-governmental organizations, and even the public at large.

Over the last five decades, the National Science Foundation (NSF) has played an important role in supporting the transformation

to data-intensive science, beginning with funding campus-based computational facilities in the 1960s, Supercomputer Center Programs in the 1980s, and the High Performance Computing and Communications program in the 1990s. In the new millennium, the Office of Cyberinfrastructure created the vision and coordinated the efforts to provide insights into complex problems in science and engineering with the help of advanced computational facilities and instruments (National Science Foundation Cyberinfrastructure Council 2007; Computer Science and Telecommunications Board, 1995).

NSF also envisioned the concept that cyberinfrastructure organizations could be created to find solutions to support data-intensive scientific and engineering research by integrating domain sciences with cyberinfrastructure, library/information sciences, and computer sciences so that data could be supported throughout its lifecycle (National Science Foundation 2007). In 2007, this was introduced as the Sustainable Digital Data Preservation and Access Network Partners, or DataNet. NSF noted that multidisciplinary approaches were needed to tackle data issues in order to (1) "provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline; (2) continuously anticipate and adapt to changes in technologies and in user needs and expectations; (3) engage at the frontiers of computer and information science and cyberinfrastructure with research and development to drive the leading edge forward; and (4) serve as component elements of an interoperable data preservation and access network" (NSF 2007).

In August 2009, NSF funded the first two DataNets -- Data Conservancy and the Data Observation Network for Earth (DataONE). This paper focuses on DataONE (http://www.dataone.org), a virtual data network focusing on the earth sciences, to explore the organization of one solution for building cyberinfrastructure and the role of librarians in that cyberinfrastructure.

Introducing DataONE

DataONE is a multi-institutional, multinational, and interdisciplinary collaboration working to develop an organizational structure that will support the full information lifecycle of biological, ecological, and environmental data and tools to be used by researchers, educators, and the public at large. DataONE focuses on enabling data-intensive biological and environmental research through cyberinfrastructure that can be used as a tool to enable new science and evidence-based policy. The key tenet is that data must be robust, accessible, and secure; therefore data management, from both the technical and sociocultural perspectives, is crucial.

> "People of all countries are experiencing increasing environmental, social, and technological challenges associated with climate variability, altered land use, population shifts, and changes in resource availability (e.g., food, water, and oil). Scientists, educators, librarians, resource managers, and the public need open, persistent, robust, and secure access to well described and easily discovered Earth observational data. Such data are critical, as they form the basis for good scientific decisions, wise management and use of resources, and informed decision-making" (Michener et al. 2009).

DataONE tackles three problems. First, DataONE provides support for studying complex environmental issues such as climate change. Environmental issues represent complex adaptive systems touching on many different disciplines. This results in studies conducted in different domains of scholarly interest (Dozier and Gaile 2009; Hunt, Baldocchi and van Ingen, 2009), making it difficult to share data and findings. An organization that serves researchers from different domains by providing a means to share data, expertise, and tools helps to bridge that gap. One example of what can be accomplished is the State of the Birds 2011 report (www.stateofthebirds.org). This is the nation's first assessment of bird distri-

bution on public lands, providing public agencies with a means to identify bird species for conservation efforts. This report was compiled from results of work done by the DataONE Scientific Exploration and Visualization Working Group.

The second problem is the lack of compatible data practices (Hunt, Baldocchi and van Ingen 2009). This problem has emerged more recently, as the value of combining the efforts of different scientists and different disciplines has been realized. Additional data challenges exist as well, including data loss (natural disaster, format obsolescence, orphaned data), scattered data sources, data deluge (the flood of increasingly heterogeneous data), poor data practices, and data longevity. An example of how DataONE is tackling this challenge is the work of the Education and Outreach Working Group, which is identifying the best practices for data curation, and producing a comprehensive, easy-to-use set of materials about best practices.

The third problem is the need to address a global problem with a global perspective (Hunt, Baldocchi and Van Ingen 2009). Many efforts have been disorganized and scattered due to disciplinary diffusion and the lack of coordination and collaboration among other stakeholders, such as governments, industry, non-governmental organizations, and citizens. There have been some successes, such as the Long Term Ecological Network (LTEeR) (http://www.lternet.edu/), that demonstrate that an organization itself can be a tool in addressing these kinds of issues.

DataONE objectives are designed to address the need for accessible, secure, and robust data, which are essential for productive research efforts and policy-making regarding environmental issues. These objectives are:

(1) providing coordinated access to current databases (such as Ecological Society for America, National Biological Information Infrastructure, Long Term Ecological Research Network and others) using the available cyberinfrastructure;

(2) creating a new global cyberinfrastructure that contains both biological and environmental data coming from different resources (e.g. research networks, environmental observatories, individual scientists, and citizen scientists);

(3) changing the science culture and institutions through the new cyberinfrastructure practices by providing education and training, engaging citizens in science, and building global communities of practice.

This leads to DataONE's mission to support science through three core areas: provision of a toolkit for data discovery, analysis, visualization and decision making; provision of easy, secure, and persistent data storage; and facilitation of community engagement of scientists, data specialists, and policy makers.

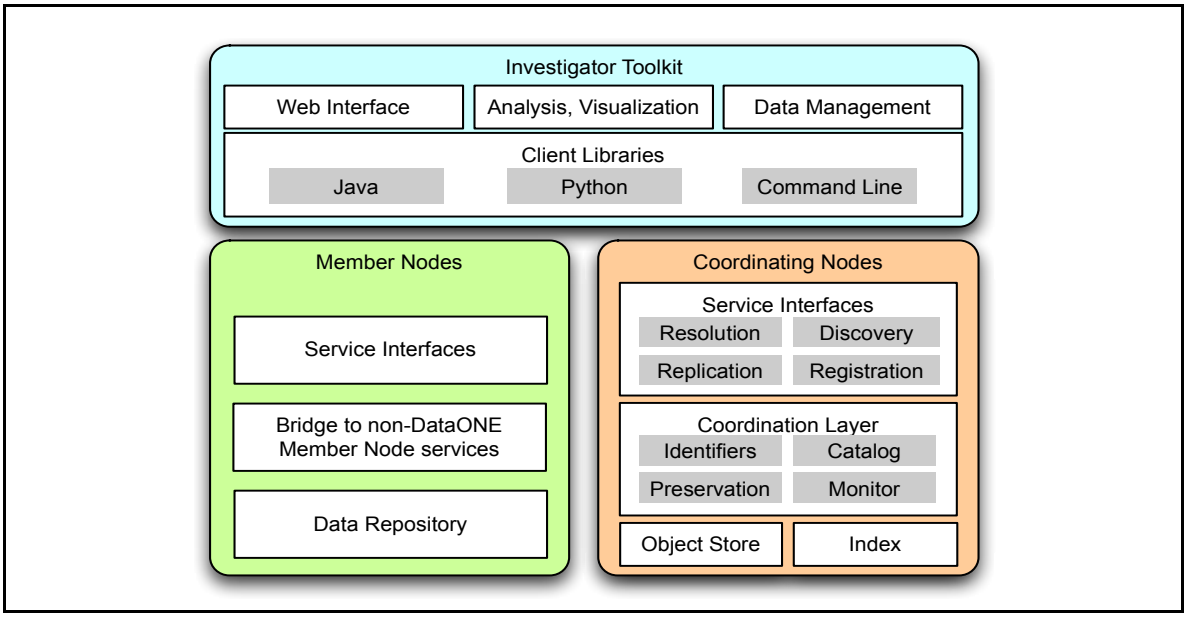
DataONE Cyberinfrastructure Primer

This paper provides only a very brief overview of the DataONE cyberinfrastructure (see Michener, et al. 2011 for more detailed information). The overall DataONE design is based on three principles. First, DataONE supports distributed management at both existing and new repositories (i.e., DataONE Member Nodes) and enables replication, caching, and discovery across these repositories for preservation, robustness, and performance. Second, the DataONE software must provide benefits for scientists and data providers today as well as adapt to tomorrow's needs. Third, DataONE activities should support and use existing community software, emphasizing free and open source software.

The cyberinfrastructure implementation of DataONE (Figure 1) is based on three major components: Member Nodes, which are ex-

Figure 1: Major Components of the DataONE Infrastructure.

Source: DataONE



isting or new data repositories that support the DataONE Member Node Application Programming Interfaces (APIs); Coordinating Nodes that are responsible for cataloging content, managing replication of content, and providing search and discovery mechanisms; and an Investigator Toolkit, which is a modular set of software and plug-ins that enables interaction with the DataONE infrastructure through commonly used analysis and data management tools.

A focus of the DataONE infrastructure is to address the problems a researcher may find when she needs content from more than one data repository, each of which may be tailored to the needs of a particular domain or community of researchers. The researcher may need to master different tools for each repository and she may need to keep separate accounts in order to access data in each of the repositories. This can be a barrier to use and may result in ambiguity as well as confusion of data authorship and access rules. The researcher in this scenario might want to retrieve content from multiple data repositories, use that content in meta-analyses or in comparison with new studies, and publish the output to a repository where others may similarly retrieve and utilize the data. DataONE architecture is developed to address the following technical challenges facing the researcher:

(1) inconsistent service interface specifications;

(2) lack of reliable unique identifier production and resolution;

(3) data longevity and availability is dependent on repository lifespan;

(4) inconsistent search semantics and effectiveness;

(5) varying service interactions and data models;

(6) access to quality metadata limits reuse of data;

(7) lack of shared identity and access control policies;

(8) difficulty in placing data near analysis,

visualization and other computational services.

DataONE and Collaboration

The DataONE cyberinfrastructure team recognizes that it is important to communicate and collaborate with others who are addressing these issues of long-term data management, reuse, discovery, and integration. There are a number of ongoing and new projects ranging from other DataNet projects to projects targeting very specific topics such as improvement of semantic search capabilities. Overlap in participation between members of the various projects helps to ensure that DataONE is up-to-date with ongoing developments and emerging approaches for data management and preservation, and also helps to ensure that other projects are aware of the base infrastructure being put into place by DataONE and how they might leverage that infrastructure.

DataONE is a Type I partner of the Federation of Earth Science Information Partners and has or is exploring collaborative relationships with many other projects including:

- other DataNet Projects,
- the Filtered-Push project (http://etaxonomy.org/mw/FilteredPush),
- the Scientific Observations Network (SONet http://www.sonet.com/),
- Semantic Tools for Ecological Data Management (SemTools https://semtools.ecoinformatics.org/),
- TeraGrid (transitioning to XD/XSEDE), and,
- the Avian Knowledge Network (AKN http://www.avianknowledge.net/content/).

DataONE's multidisciplinary environment requires vibrant collaboration in order to pursue the organizational goal stated on the website: "DataONE will be commonly used by researchers, educators, and the public to better understand and conserve life on earth and the environment that sustains it." The organization is built around environmental scientists, with a strong collaboration with information scientists. Each of these groups is highly diversified. The environmental sciences include scientists from biology, ecology, environmental sciences, hydrology, and biodiversity. The information science members include specialists in informatics, computer engineering, computer sciences, information sciences, information management, information technology, and library sciences.

In the future, DataONE envisions ever-strengthening collaborations involving more associated disciplines. For instance, possible areas for expansion include researchers studying migration and urbanization, such as sociologists, and those studying natural resource allocation, such as economists. DataONE's goal, and challenge, is to create the cyberinfrastructure that can address multi-faceted environmental issues and mobilize all the interested parties to engage.

DataONE is also highly collaborative in terms of institutions. At DataONE's inception in August 2009, DataONE partners included Cornell University, the National Evolutionary Center at Duke University, Oak Ridge National Laboratory, the University of New Mexico, the California Digital Library at the University of California, the National Center for Ecological Analysis and Synthesis at the University of California Santa Barbara, the University of Illinois-Chicago, The University of Tennessee-Knoxville, the University of Kansas, the U.S. Geological Survey, and Utah State University. The diversity of these initial institutions can be seen in Table 1. This list of partners continues to grow.

As noted in the earlier section, the technological design creates collaboration at two levels of participation: Coordinating Nodes (the initial ones are the University of New Mexico, the partnership between University of Tennessee and Oak Ridge National Laboratories, and the National Center for Ecological Analysis and Synthesis at the University of California, Santa Barbara) and Mem-

Table 1: Institutions that are involved in or are supporting DataONE activities on different levels

Academic institutions from the U.S. (including three EPSCoR [The Experimental Program to Stimulate Competitive Research] states—Tennessee, Kansas, and New Mexico) and the United Kingdom (i.e., Edinburgh, Manchester, Southampton);
Research networks (e.g., Long Term Ecological Research Network, Consortium of Universities for the Advancement of Hydrologic Science Inc. [CUAHSI], Taiwan Ecological Research Network, South African Environmental Research Network [SAEON]);
Environmental observatories (e.g., The National Ecological Observatory Network [NEON], USA-National Phenology Network, Ocean Observatory Initiative, South African Environmental Observatory Network);
NSF- and government-funded synthesis (i.e., the National Center for Ecological Analysis and Synthesis [NCEAS], the National Evolutionary Synthesis Center [NESCent], Atlas of Living Australia) and supercomputer centers/networks (Oak Ridge National Laboratories [ORNL], National Center for Supercomputing Applications [NCSA], and TeraGrid);
Governmental organizations (e.g., U.S. Geological Survey [USGS], the National Aeronautics and Space Administration [NASA], Environmental Protection Agency [EPA]);
Academic libraries (e.g., University of California Digital Library, University of Tennessee, and University of Illinois-Chicago libraries, which are active in the digital library community and are members of the Coalition for Networked Information, the Digital Library Federation, and the Association of Research Libraries);
International organizations (e.g., Global Biodiversity Information Facility, Inter American Biodiversity Information Network, Biodiversity Information Standards);
Numerous large data and metadata archives (e.g., USGS-National Biological Information Infrastructure, ORNL Distributed Active Archive Center for Biogeochemical Dynamics, World Data Center for Biodiversity and Ecology, Knowledge Network for Biocomplexity);
Professional societies (e.g., Ecological Society of America, Natural Science Collections Alliance);
NGOs (e.g., The Keystone Center); and
The commercial sector (e.g., Amazon, Battelle Ventures, IBM, Intel)
Source: DataONE Proposal, 2009.

ber Nodes (the first three are in the process of coming online as of this writing). Coordinating Nodes are geographically-distributed to provide a high-availability, fault-tolerant, and scalable set of coordinating services to the Member Nodes. They are responsible for utility services across the collaboration: member node registration services, metadata indexing, coordinating and monitoring data replication, providing global user identity services, providing log aggregation services, and monitoring node and network health. Member Nodes will be located inside academia, libraries, government agencies, and other organizations to provide local data storage, data access, access control, replication, metadata quality, and primary user interaction.

DataONE Organizational Structure

DataONE's organizational structure includes a small managerial team (principal investigator, executive director, and directors for cyberinfrastructure and community engagement), as well as a core cyberinfrastructure team that is responsible for designing and building the cyberinfrastructure. The Management Team is based at the University of New Mexico with principal investigator Dr. William Michener, who is Professor and Director of eScience Initiatives at the University Libraries at the University of New Mexico.

While DataONE is domain-centric and led by domain scientists, librarians and information scientists are integral members of the team at all levels and across many activities. For example, the Leadership Team, which meets each week in a virtual environment to share and coordinate technical and sociocultural activities, is composed of 14 individuals (in addition to the management team) representing 10 institutions, five of whom are librarians or information scientists. DataONE is also advised by two external bodies – the External Advisory Board and the DataONE Users Group, each of which has librarian and information science representation.

DataONE is a virtual organization with a strong network of people. The network is built around working groups that help assure that multiple perspectives are represented. Working groups are composed of scientists, academic researchers, educators, government and industry representatives, and leading computer, information, and library scientists. Working groups are central to DataONE research activities and most focus on either cyberinfrastructure or community engagement issues, although two working groups, Usability and Assessment and Exploration, Visualization, Analysis, directly engage in both cyberinfrastructure and community engagement activities. The Working group model allows DataONE to conduct targeted research and education activities with a broad group of scientists and users. Working groups are also designed to enable research and education activities to evolve over time.

Each working group has two co-leaders, at least one of whom is a member of the Leadership Team, in order to facilitate communication between groups and to assure that the Management Team is aware of all activities. Each group has an additional 8-10 members who are actively engaged in ongoing activities. Also, there is periodic interaction among the working groups such as members of different working groups addressing a problem together. When there are face-to-face meetings, there are sessions devoted to join forces and perspectives. The structure is fluid, flexible, and adaptive.

DataONE Lifecycle

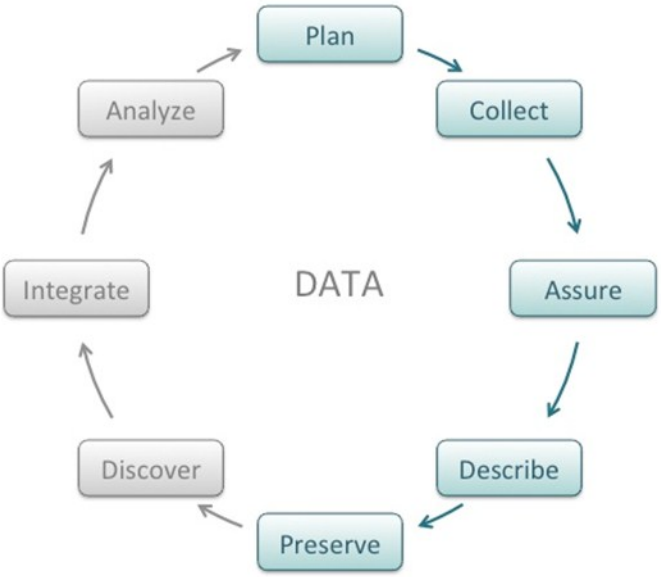
Through the activities of working groups, DataONE addresses the complete data lifecycle through a comprehensive program of research, design, and development to create a system to preserve, disseminate, and protect research objects in a secure, reliable, and open approach that is responsive to users' and scientists' needs.

DataONE has adopted a lifecycle model that focuses on "the data" and illustrates the different stages that data can pass through, although data may skip a stage or stages (Michener et al. 2011). At each stage, different people may interact with the data, and it is unlikely that one person will interact with the data at all stages. The data lifecycle is useful because it can be used to identify dataflows and work processes for scientists, librarians, or others associated with the science data process.

Let's follow the data through the eight stages of the lifecycle (Figure 2). The lifecycle begins when scientists make a plan to conduct their research. They then collect data either in the field or laboratory. The scientific team may then review the data to assure the data

Figure 2: DataONE has adopted a data lifecycle which focuses on the way data moves through eight unique stages. These steps begin at the point of creating the research plan then progress to data collection, quality assurance and quality control. Data needs to be described – which is when metadata is created. Data are then deposited in a trusted repository where they may be preserved. Tools and services can then support data discovery, integration, and analysis including visualization.

Source: DataONE

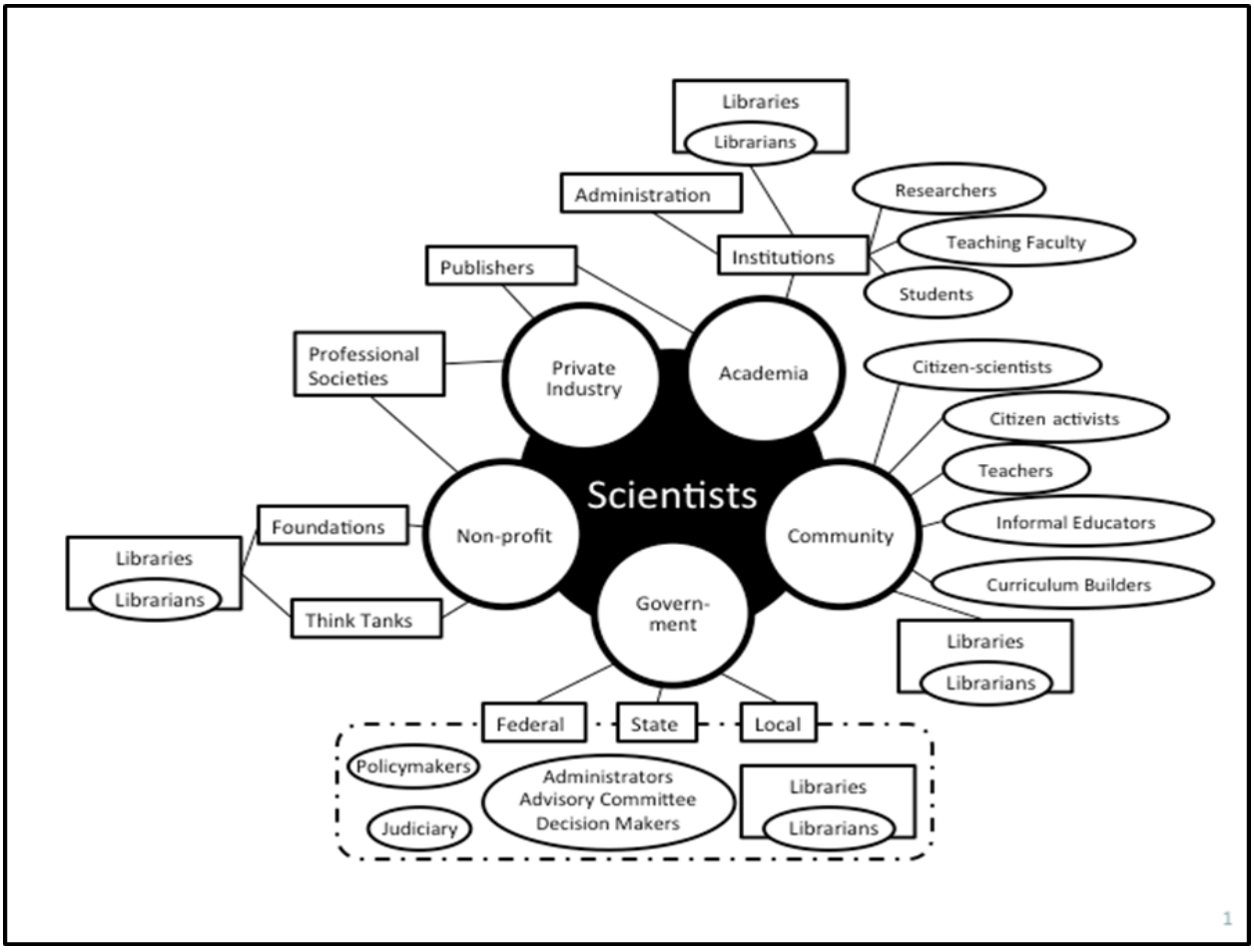


quality. The data is now ready to be described with metadata. Although it is recommended that the specific domain metadata standard be used, scientists often use a metadata schema that has been developed for their project. When the data are described, they are ready to be deposited into a trusted repository in which they will be preserved. The data are now discoverable and may be accessed by others. At this point, data modelers or other scientists might access the data and integrate multiple data sets for analysis. Conversely, data may not be integrated and may instead be analyzed by the original scientist who collected it (skipping both the discover and integrate stages).

Librarians can provide support and guidance at nearly every stage of the data lifecycle. At the Planning stage, librarians can address data management questions that can help scientists develop a data management plan. In 2011, NSF began requiring that a data management plan be submitted with each proposal. Librarians have played a very active role during the development of a new tool, the DMPTool (https://dmp.cdlib.org/), that helps researchers create data plans online. The DMPTool original partner institutions include four libraries and the United Kingdom's Digital Curation Centre. At the Assure stage, librarians can help scientists identify existing strategies for data quality. At the Describe stage, librarians can help the scientist identify and apply a rele-

Figure 3: DataONE stakeholders. Scientists are the primary stakeholders and circles represent each of five science research environments. There are secondary stakeholders associated with each science research environment. Organizations are represented with boxes and individuals with ovals. The dashed box indicates stakeholders associated with each level of government.

Source: Michener, et al., 2011



vant metadata schema. At the Preserve stage, librarians can identify or perhaps their library can provide an appropriate and trustworthy repository. At the Discover stage, librarians can help users find and access data, which is an extension of a traditional role of librarianship. Librarian engagement at the Integrate stage is still evolving, however it may include helping to negotiate the interconnected information challenges noted earlier.

DataONE Stakeholders

DataONE engages a wide group of stakeholder communities (Figure 3). The primary stakeholders at the center of the stakeholder network are scientists. Scientist practices and attitudes vary depending on the home domain, meaning that scientists are not homogenous. Science communities were not categorized by domain since that approach discouraged crossing disciplinary boundaries and practicing integrative science. Rather, this stakeholder community was characterized based on how scientists “do” science

that aligned with five *science research environments*: academia, government, private industry, non-profit, and community. While scientists working in private industry are primary stakeholders, the proprietary constraints placed on them means they are likely to be restricted from sharing and therefore may have a limited relationship with DataONE.

Secondary stakeholders also have a role in data-intensive environmental science. The following are the major groups of secondary stakeholders:

(1) Libraries and librarians are important sources of support for science and scientists to negotiate the data-driven and information-reliant milieu in any of the five science research environments. DataONE prioritizes libraries and librarians as the most important secondary community. In the DataONE stakeholder network, the definition of libraries includes the full range of information-centric agencies and services;

(2) Administrators and policy makers at the federal, state, and local level are people influencing the success of science through funding programs and policy that may facilitate or hinder research;

(3) Publishers and professional societies whose activities include the dissemination of research results and data;

(4) Think tanks which develop evidence-based position papers or policy suggestions;

(5) Citizen scientists, citizen activists, K-12 teachers, informal educators, and curriculum builders. These stakeholders provide the bridge between science and the public.

Libraries, Librarians & DataONE

From the proposal stage, DataONE had library and information science (LIS) professionals and researchers on the team. This has provided the library and information center perspective as the cyberinfrastructure has emerged. As the cyberinfrastructure matures, LIS professionals and researchers serve as on-going members of the leadership team and working groups helping to shape how DataONE addresses the issues and builds technical and sociocultural infrastructure. LIS professionals have the experience and knowledge to help understand how stakeholders interact across the five science research environments and throughout the data lifecycle. This section focuses primarily on the sociocultural contributions, although there are also LIS professionals engaged in answering DataONE's technical questions.

In sociocultural terms, LIS skills and tools are helping provide insight into stakeholders' motivation, practices, and needs. This is being accomplished through a series of assessments being conducted with different DataONE stakeholder communities. These assessments are designed to explore attitudes towards, and practices for, science data. The results help developers have a better understanding of how these targeted communities are engaging with science data and help developers create tools that will provide useful services and also be usable by the community. For instance, LIS research shows that scientists' data practices (data management, digital curation, metadata creation, and data preservation) are poor for various reasons, including a lack of knowledge of existing tools and a lack of desire to use them (Tenopir et al. 2011; Parse Insight 2010).

Research conducted to learn more about this include surveys, interviews, usability studies, and analyses of data use. Working group efforts have been instrumental in conducting baseline assessments with stakeholders, analyzing the assessment studies that others have done, and conducting repeat assessments of various stakeholder groups every couple of years. These as-

sessments help identify current data needs, perceptions, and practices of all parts of the data lifecycle and provides a basis for seeing how these change over time. The DataONE baseline assessment of scientists was completed last year (Tenopir et al. 2011). Baseline assessments are now being conducted with libraries, librarians, and data managers.

Librarians have been actively involved in the teams, creating "user scenarios" for primary stakeholder groups. These exemplify how a user might interact with DataONE and highlight specific activities that they would be engaged in. By understanding the stakeholders' needs, motivations, concerns, and skill base, DataONE developers can better develop appropriate tools and services, and also understand the best way to market them effectively to appropriate groups.

Librarians have also been key members on the team developing personas. A persona provides a way to envision the "average" user in a particular stakeholder group. The personas were developed using data from the assessment surveys and from interviews. The personas help build an understanding of current and potential users. Personas allow developers, LIS professionals, and DataONE management to visualize how users from specific communities may use DataONE. Knowing this facilitates building better tools and providing better services, and also increases the ability to make good strategic decisions as the cyberinfrastructure grows.

Librarians are also serving on working groups that are responsible for the development of a large number of best practices for data management, and for maintaining a list of tools available for a range of data activities, including visualization and management (http://www.dataone.org/dataonepedia). Educational modules for data management are also being developed. Many of these resources can be found at http://www.dataone.org/resources.

Conclusion

The data-intensive environment is changing the way scientists "do" science, and as librarians, we can provide support and services that will help scientists concentrate on "doing" science rather than wrestling with barriers that keep them from creating sharable datasets and from utilizing the range of data available. At universities across the country, libraries are facing questions about how to address eScience challenges. Librarians are assessing what skills they need to provide the services associated with eScience, as well as how to integrate these new responsibilities into their workload. DataONE provides a laboratory to address the new roles and responsibilities facing science librarianship.

The DataONE data lifecycle helps identify some of the areas where librarians can offer essential skills and support. Librarians are important partners from the moment scientists begin planning their data collection by providing information and support for creating data management plans. There are also roles in this dynamic new information environment that are based on the very foundations of librarianship: metadata creation, preservation strategies, and information access. It is likely that these activities have a very different look in the eScience context; however, the basic tenets established from years of research and practice provide a strong foundation for developing the specialized skill set.

Another area that DataONE illuminates is the successful partnering of librarians and information science researchers with domain scientists. There is much potential for librarians to become more integrated in the science workflow. This includes working closely with scientists at all stages of the data lifecycle, as well as participating in the data literacy education of the next generation of scientists by helping coach science undergraduates and graduates on best practices related to data.

EScience presents a host of challenges for libraries, including having sufficient technical capacity, a workforce with the training to address eScience, and the capacity to do the outreach and training needed to engage scientists and student scientists. Committing resources at the library level can be difficult without strong institutional support, especially since science researchers may not envision how their data management can benefit from greater interaction with the library. This lack of recognition may make it more difficult for the library to illustrate the return on investment. The DataONE experience suggests that libraries can be involved in high profile activities such as creating data management plans, providing metadata guidance, and identifying reliable data repositories. Since these activities protect university intellectual assets, they may help establish the value of supporting library involvement with eScience.

Librarians also face the challenge of finding ways to integrate eScience activities into their work day. The DataONE experience suggests that librarians are invaluable partners in data description, preservation and access. The necessary skill set is based on librarianship fundamentals, but does require the librarian to become acquainted with specific best data practices. Many associations are offering workshop opportunities, but librarians may also utilize resources such as the DataONE best practices and tools archives.

Libraries and librarians have a history of successfully adjusting to a shifting information landscape. As evidenced by librarian participation in DataONE, the library community is already an active partner in shaping the future of eScience.

References

Computer Science and Telecommunications Board. 1995. *Evolving the high performance computing and communications initiative to support the nation's information infrastructure*. Washington, D.C.: The National Academies Press.

Dozier, Jeff and William B. Gail. 2009. "The Emerging Science of Environmental Applications." In *The Fourth Paradigm: Data-Intensive Scientific Discovery,* edited by Tony Hey, Stewart Tansley, and Kristin Tolle, 13-19. Microsoft Research. Accessed June 29, 2011. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

Gray, Jim. 2007. "eScience Talk." Talk presented to the NRC-CSTC, Mountain View, CA, January 11.

Hannay, Timo. 2009. "From Web 2.0 to the Global Database." In *The Fourth Paradigm: Data-intensive Scientific Discovery,* edited by Tony Hey, Stewart Tansley, and Kristin Tolle, 215-220. Microsoft Research. Accessed June 29, 2011. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

Hey, Tony, Stewart Tansley, and Kristin Tolle, eds. 2009. *The Fourth Paradigm: Data-intensive Scientific Discovery.* Microsoft Research. Accessed June 29, 2011. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

Hunt, James R., Dennis D Baldocchi, and Catharine van Ingen. 2009. "Redefining Ecological Science Using Data." In *The Fourth Paradigm: Data-intensive Scientific Discovery,* edited by Tony Hey, Stewart Tansley, and Kristin Tolle, 21-26. Microsoft Research. Accessed June 29, 2011. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

International Panel on Climate Change. 2007. *Climate change 2007: Synthesis report.* Retrieved on June 30, 2011 from http://

www.ipcc.ch/pdf/assessment-report/ar4/syr/ar4_syr.pdf

Kafel, Donna. 2010. "e-Science and its Relevance for Research Libraries." Accessed July 30, 2011. http://esciencelibrary.umassmed.edu/escience

Lynch, Clifford. 2009. "Jim Gray's Fourth Paradigm and the Construction of the Scientific Record." In *The Fourth Paradigm: Data-intensive Scientific Discovery,* edited by Tony Hey, Stewart Tansley, and Kristin Tolle, 177-183. Microsoft Research. Accessed June 29, 2011. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

Michener, William, Todd Vision, Stephanie Hampton, and Robert Cook. 2009. "DataONE Proposal."

Michener, William K., Suzie Allard, Amber Budden, Robert Cook, Kimberly Douglass, Mike Frame, Steve Kelling, Rebecca Koskela, Carol Tenopir, and David A. Vieglais. 2011. "Participatory Design of DataONE - Enabling Cyberinfrastructure for the Biological and Environmental Sciences." *Ecological Informatics.* Accepted, in press, available online 3 September 2011. Accessed 5 September 2011. http://dx.doi.org/10.1016/j.ecoinf.2011.08.007

National Academies of Science. Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. 2009. "Ensuring the integrity, accessibility, and stewardship of research data in the digital age." Accessed August 2, 2011. http://www.nap.edu/catalog.php?record_id=12615

National Science Foundation. 2007. "Sustainable Digital Data Preservation and Access Network Partners (DataNet)." Accessed July 26, 2011. http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm

National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. 2003. *Revolutionizing Science and Engineering through Cyberinfrastructure.* Accessed on August 20, 2011. www.nsf.gov/od/oci/reports/atkins.pdf

National Science Foundation Cyberinfrastructure Council. 2007. *Cyberinfrastructure Vision for 21st Century Discovery*. Accessed July 29, 2011. http://www.arl.org/bm~doc/ci_vision_march07.pdf

PARSE Insight. 2010. *PARSE Insight.* Accessed January 31, 2011. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff and Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions." *PLoS One* 6 (6):e21101 (2011) doi:10.1371/journal.pone.0021101

World Energy Council. 2010. *Energy and urban innovation*. Accessed February 2, 2011. http://www.worldenergy.org/documents/eui_2010_1.pdf

World Health Organization. 2009. *World now at the start of 2009 influenza pandemic.* Accessed August 30, 2011. http://who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/index.html

Disclosure: The author reports no conflicts of interest.

ISSN 2161-3974